UNIVERSITEIT GENT

Studiefiche

Vanaf academiejaar 2020-2021

Gedistribueerde gegevensverwerking (E761040)

Wegens Covid19 kan mogelijk afgeweken worden van de onderwijs- en evaluatievormen. Dergelijke afwijkingen zullen via Ufora worden gecommuniceerd.

Cursusomvang (nominale waarden; effectieve waarden kunnen verschillen per opleiding)

Studiepunten 3.0 Studietijd 90 u Contacturen 30.0 u

Aanbodsessies in academiejaar 2020-2021

A (semester 1) Nederlands Gent

Lesgevers in academiejaar 2020-2021

Volckaert, Bruno	TW05	Verantwoordelijk lesgever

Aangeboden in onderstaande opleidingen in 2020-2021	stptn	aanbodsessie
Bachelor of Science in de industriële wetenschappen (afstudeerrichting informatica)	3	A
Master of Science in de industriële wetenschappen: informatica	3	A
Uitwisselingsprogramma industriële wetenschappen: informatica	3	A

Onderwijstalen

Nederlands

Trefwoorden

gedistribueerde gegevensopslag, parallelle gegevensverwerking, dataverwerkingsmodellen, Big Data, Business Intelligence, Computerwetenschappen, Informatica, Computertechnologie

Situering

In dit opleidingsonderdeel worden de uitdagingen en technieken bestudeerd om business intelligence te halen uit grote volumes data en/of een veelheid aan datastromen. Hiertoe wordt een “data pipeline” opgezet: het verzamelen en extraheren van data uit interne en externe bronnen, het opslaan van deze ruwe data, het omvormen naar kwalitatieve en bruikbare data en tot slot het analyseren en visualiseren van de data.
Studenten verwerven inzicht in parallele gegevensverwerking, zowel vanuit een computationeel-uitvoerend standpunt (onderliggende werking van hedendaagse platformen) als vanuit programmeertechnisch-algoritmisch standpunt.

Inhoud

De computationeel-uitvoerende concepten die aan bod komen zijn:
• Data verzamelen uit verschillende bronnen (bestanden, SQL, NoSQL, streaming data, scraping)
• Gedistribueerde bestandssystemen als grondslag voor een data lake
• Programmeermodellen en architecturen voor het verwerken van data
• Lambda, Kappa en Zeta architectuur
• batch-processing, mini-batch, streaming
• MapReduce en hedendaagse varianten
Onderliggende mechanismes voor parallelle gegevensverwerking (“Bringing compute to the data”) op een gedistribueerd bestandssysteem
• Redundantie
• Sorteren en samenbrengen
• Parallelle uitvoering
• Identificatie van mogelijke flessenhalzen
De programmeertechnische concepten die aan bod komen zijn de volgende:
• Het omzetten van ruwe data naar kwalitatieve data (data cleaning)
• Ruis, normalisatie, transformeren naar geschikt formaat
• Exploratieve data-analyse
• Organisatie van datasets
• (Re)-sampling

Data warehouse (vs data lake) met gestructureerde data voor Business Intelligence
- Rol van data management: metadata, data discovery, data ownership, data privacy
- Data warehouse vs data mart vs operational data store
- Extract-Transform-Load principe
- Facts, dimensions, star schemas, schema evolution ...
- Online Analytical Data Processing databases geoptimaliseerd voor BI-workloads: weinig gebruikers, veel data

Mogelijke technologieën die de bovenstaande concepten ondersteunen zijn
- Extractie van data uit bronnen: Sqoop, Flume, Hadoop Connector
- Exploratieve data-analyse: Python Pandas
- Data lake: Hadoop DFS en Yarn
- Configureren van data workflows: Apache Spark Streaming, Apache Hive, Apache Pig, Apache Beam, Apache Flink, etc.
- Notebooks en visualisatie van data: Tableau, Plotly, Jupyter notebook, Apache Zeppelin

Begincompetenties

Objectgericht programmeren en softwareontwikkeling, Python

Eindcompetenties

1 Studenten kunnen grote en gevarieerde data analyseren met behulp van gedistribueerde gegevensverwerking
2 De studenten kennen de programmeermodellen en platformen voor verwerking van grote en gevarieerde data
3 De studenten kennen de technieken en visualisaties van exploratieve data-analyse

Creditcontractvoorwaarde

Toelating tot dit opleidingsonderdeel via creditcontract is mogelijk mits gunstige beoordeling van de competenties

Examencontractvoorwaarde

Dit opleidingsonderdeel kan niet via examencontract gevolgd worden

Didactische werkvormen

Online groepswerk, online hoorcollege, online werkcollege

Toelichtingen bij de didactische werkvormen

Hoorcollege, werkcollege: oefeningen, groepswerk (project)

Leermateriaal

Slides op de elektronische leeromgeving

Referenties

Vakinhoudelijke studiebegeleiding

Interactieve ondersteuning via het elektronische leerplatform; begeleide project-oefeningen; contact met de lesgever en assistenten via e-mail en persoonlijk na afspraak.

Evaluatiemomenten

periodegebonden en niet-periodegebonden evaluatie

Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode

Schriftelijk examen met open vragen

Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode

Schriftelijk examen met open vragen

Evaluatievormen bij niet-periodegebonden evaluatie

Werkstuk, verslag

Tweede examenkans in geval van niet-periodegebonden evaluatie

Examen in de tweede examenperiode is mogelijk

Toelichtingen bij de evaluatievormen

Eerste examenkans:
- PE1: schriftelijk examen met open vragen
- NPE1: beoordeling van eindresultaat project op basis van verslag / code

Tweede examenkans:
- PE2: schriftelijk examen met open vragen
- NPE2: beoordeling van eindresultaat project op basis van verslag / code

Eindscoreberekening

• 50% van het eindcijfer wordt bepaald door de antwoorden op het schriftelijk examen
• 50% van het eindcijfer wordt bepaald door de evaluatie van het project
Om te kunnen slagen voor het opleidingsonderdeel moet minstens 9/20 behaald worden voor PE. Is aan deze voorwaarde niet voldaan, dan wordt er afgeweken van het berekende cijfer indien dit 10 of meer is en haalt de student een 9/20.